EFFICIENT RECURSIVE CLUSTERING BASED ON A SPLITTING FUNCTION DERIVED FROM SUCCESSIVE EIGEN-DECOMPOSITIONS

Field of the Invention

The present invention relates generally to data clustering, and to methods and arrangements for facilitating the same, such as in the context of enrolling target speakers in a speaker verification system. The present invention relates more particularly to the partitioning of a set of multidimensional data points into classes.

Background of the Invention

It is generally desired that the classes, when modeled with Gaussian densities for example, can be used to construct a probability density for the data. Additional data obtained in the same way as the original set should be judged highly likely according to the constructed density. Clustering is a fundamental data analysis tool and is the basis for many approaches to pattern recognition. Among other things, this process facilitates analyzing the areas of the data space that are the most concentrated with points, while allowing one to determine which points may be outliers (i.e., data points that result from noise and do not give information about the process or system being modeled). It also forms the basis for a compact representation of the data.

Clustering is usually a very time consuming process requiring many iterative passes over the data. Generally, the clustering problem is handled by a clustering technique such as K-means or LBG (see Y. Linde, A. Buzo, R. M. Gray, "An Algorithm for Vector Quantizer Design," IEEE Trans. Commun., vol. 28, pp.84-95, Jan. 1980). K-means starts with an initial seed of classes and iteratively re-clusters and re-estimates the centroids. The effectiveness of this method depends on the quality of the seed. LBG does not require a seed, but starts with one cluster for all of the data. Then, it uses a random criterion to generate new centroids based on the current set (initially one). K-means is used after constructing the new set of centroids. The process is repeated on the new set. In K-means, the requirement for a good seed is strong, which means one needs a lot of prior information. The iterative reclusterings are also time consuming. LBG has a random component which makes it potentially unstable in the sense that quite different models can result from two independent LBG clusterings of the same data.

In view of the foregoing, a need has been recognized in connection with improving upon the shortcomings and disadvantages associated with conventional data clustering methods and arrangements.

Summary of the Invention

In accordance with at least one presently preferred embodiment of the present invention, clustering problems are solved in an efficient, deterministic manner with a recursive procedure to be discussed below.

In summary, the present invention provides, in one aspect, an apparatus for facilitating data clustering, the apparatus comprising: an arrangement for obtaining input data; and an arrangement for creating a predetermined number of non-overlapping subsets of the input data; the arrangement for creating a predetermined number of non-overlapping subsets being adapted to split the input data recursively.

In another, aspect, the present invention provides a method of facilitating data clustering, the method comprising the steps of: obtaining input data; and creating a predetermined number of non-overlapping subsets of the input data; step of creating a predetermined number of non-overlapping subsets comprising splitting the input data recursively.

Furthermore, the present invention provides, in an additional aspect, a program storage device readable by machine, tangibly embodying a program of instructions executable by the machine to perform method steps for facilitating data clustering, the method comprising the steps of: obtaining input data; and creating a predetermined

number of non-overlapping subsets of the input data; step of creating a predetermined number of non-overlapping subsets comprising splitting the input data recursively.

For a better understanding of the present invention, together with other and further features and advantages thereof, reference is made to the following description, taken in conjunction with the accompanying drawings, and the scope of the invention will be pointed out in the appended claims.

Brief Description of the Drawings

Fig. 1 schematically illustrates a general clustering operation.

Fig. 2 schematically illustrates a splitting procedure.

Fig. 3 schematically illustrates a tree of splits.

Description of the Preferred Embodiments

A recursive procedure for solving clustering problems, in accordance with at least one presently preferred embodiment of the present invention, is discussed herebelow.

With reference to Figure 1, Let $\mathbf{X} = \{x_i\}$ be the set of points 101 to cluster in a real n dimensional space. Thus each $\mathbf{x}_i$ is a vector of n real numbers.

$$x_i = \{r_1, r_2, \ldots, r$$

As shown in Figure 1, the goal of the clustering process 102 is generally to create N non-overlapping subsets $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ of the original data $\mathbf{X}$ (indicated at 104). To this end, the following splitting procedure (with reference to Figure 2) is preferably applied recursively to the data 104in accordance with an embodiment of the present invention:

- Define $\mathbf{X}_a$ (108) to be the input and $\mathbf{X}_b$ and $\mathbf{X}_c$ to be the outputs (118/120).

- Compute (109) $\mathbf{m}_a$ = the mean of $\mathbf{X}_a$ and for every x_i in $\mathbf{X}_a$, subtract out this mean. Thus the input data will now be "zero mean". Denote the resulting set of vectors by $\mathbf{X}_a'$.

- Compute the eigenvector decomposition of $\mathbf{X}_a'$ (110) and let $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ be the eigenvectors arranged in order of decreasing eigenvalue magnitudes.

- For each vector x_i' in $\mathbf{X}_a'$, compute (112) the vector of projection coefficients onto the set of eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$:

$$x_i' \;\rightarrow\; \{\langle x_i', e_1\rangle, \langle x_i', e_2\rangle, \ldots, \langle x_i', e_n\rangle\} \;=\; \{c_1, c_2, .$$

where <,> indicates dot product.

- Let $f(\{c_1, c_2, ..., c_n\})$ be a (non)linear function of the projection coefficients with real valued output (indicated at 114).

- Let d_f be the probability density of f (indicated at 116).

- Define t_0 to be the value of f for which $d_f = 1/2$. This is indicated as "threshold" (118) in Fig. 2. (Though here a "2-way" split of data is contemplated with a threshold of 1/2, it should be understood that other and/or more thresholds could be chosen to split the data into essentially any number of N subsets with N-1 thresholds. Thus, for example, if a "3-way" split of data is contemplated, then thresholds could be chosen at $d_f = 1/3$ and $d_f = 2/3$. Accordingly, for N subsets, there will be N-1 thresholds whose values are m/N, with m being an integer number from 1 to N-1.)

- Define $\mathbf{X}_b$ and $\mathbf{X}_c$ (118/120) as follows:

$$X_b = \{x_i : f(\{c_1, c_2, ..., c_n\}_i)$$

$$X_c = \{x_i : f(\{c_1, c_2, ..., c_n\}_i)$$

The clustering is preferably initialized with $\mathbf{X}_a = \mathbf{X}$. After $\mathbf{X}_a$ is split into $\mathbf{X}_b$ and $\mathbf{X}_c$, each of these in turn becomes the input to the procedure. Thus $\mathbf{X}_b$ and $\mathbf{X}_c$ are split in the same way each into two subsets. The procedure is repeated until the desired number

of subsets are created. These splitting procedures are illustrated schematically in Figure 3. (The individual splitting procedures ["SP"] are indicated at 122.)

Among the advantages of the method presented hereinabove are the following:

- There is no variability due to randomness, as the clustering is completely deterministic. This is important when it is required that multiple clusterings of identical data result in the same final classes.

- The splitting procedure seeks maximum separability, based on an eigenvector analysis.

- There is no need for an initial seed, such as (for example) the one needed in K-means clustering.

- Greater speed results. Since there is no requirement for an iterative reclustering procedure, the method can be performed very quickly. Experimentation has shown that, in comparison with LGB, performance is never compromised and, in fact, is often enhanced.

- Each node (124) in the tree shown in Figure 3 represents a class of data. Statistical models may now (once all of the splitting is complete) be built for each of these classes using any desired technique. The resulting models can be used for pattern classification.

A practical application of the techniques discussed and contemplated herein is in the enrollment of target speakers in a speaker verification system. In this case, if it is desired that models be built as quickly as possible, then the techniques described and contemplated herein can speed up training time by a significant order of magnitude. An example of a speaker verification system that may readily employ the embodiments of the present invention is discussed in U. V. Chaudhari, J. Navratil, S. H. Maes, and Ramesh Gopinath, "Transformation Enhanced Multi-Grained Modeling for Text-Independent Speaker Recognition", ICSLP 2000, pp. II.298-II.301.

It is to be understood that the present invention, in accordance with at least one presently preferred embodiment, includes an arrangement for obtaining input data; and an arrangement for creating a predetermined number of non-overlapping subsets of the input data, which together may be implemented on at least one general-purpose computer running suitable software programs. These may also be implemented on at least one Integrated Circuit or part of at least one Integrated Circuit. Thus, it is to be understood that the invention may be implemented in hardware, software, or a combination of both.

If not otherwise stated herein , it is to be assumed that all patents, patent applications, patent publications and other publications (including web-based

publications) mentioned and cited herein are hereby fully incorporated by reference herein as if set forth in their entirety herein.

Although illustrative embodiments of the present invention have been described herein with reference to the accompanying drawings, it is to be understood that the invention is not limited to those precise embodiments, and that various other changes and modifications may be affected therein by one skilled in the art without departing from the scope or spirit of the invention.